Students' perception of school moral atmosphere: From moral culture to social competence. A generalizability study.

Beem, A.L.; Brugman, D.; Host, K.; Tavecchio, L.

published in
European Journal of Developmental Psychology
2004

DOI (link to publisher)
10.1080/17405620444000076

document version
Publisher's PDF, also known as Version of record

EUROPEAN JOURNAL OF DEVELOPMENTAL PSYCHOLOGY, 2004, *1*(2), 171–192

Students' perception of school moral atmosphere: From moral culture to social competence. A generalizability study

A. Leo Beem
Department of Biological Psychology, Vrije Universiteit, The Netherlands

Daniel Brugman
Department of Psychology, Universiteit Utrecht, The Netherlands

Karin Høst
Lauwerbes 5, NL-2318 AT Leiden, The Netherlands

Louis W. C. Tavecchio
Center for Child and Family Studies, Universiteit Leiden, The Netherlands

The School Moral Atmosphere Questionnaire (SMAQ) was constructed to measure differences in students' perception of school moral atmosphere between schools. The instrument is based upon the constructs defined by the Just Community Approach that focused on students' shared perspective and portrayed ideal types of school moral culture. This study presents reliability estimates of the SMAQ based on a generalizability study. A total of 1280 students from 32 normal secondary schools participated in the study. The design of the study includes the factors: type of school varying in educational level; school; grade level; class; and student. Variance components and reliabilities are estimated for two models. In Model 1 grade level is a fixed effect, in Model 2 grade level is a random effect. The results indicate that moral atmosphere in school can be measured reliably, although in Model 2 a considerable number of observations may be needed. Because score levels for some subscales depend on the school type, reliabilities are higher for the entire population than for populations consisting of one particular school type. It is concluded that students' perception of moral atmosphere in normal secondary schools have a strong individual flavour. Perceived moral atmosphere should

Please address all correspondence to Daniel Brugman, Department of Psychology, Universiteit Utrecht, Heidelberglaan 2, NL-3584 CS Utrecht, The Netherlands. E-mail: D. Brugman@FSS. UU. NL

The authors are indebted to the Netherlands Organization for Scientific Research (NWO) for its financial support of project 590–290–501. The second author is also indebted to NWO for grant no R 56–460.

http://www.tandf.co.uk/journals/pp/17405629.html DOI: 10.1080/17405620444000076

not be regarded primarily as a shared perspective among students within a school like a moral school culture, but as an instance of the social competence of the individual student.

Until quite recently, studies on moral development were primarily concerned with moral competence: the highest level of moral judgement that individuals achieve when they are asked to reason about abstract hypothetical moral dilemmas. These studies grew out of Kohlberg's developmental theory of moral judgement (Colby & Kohlberg, 1987; Kohlberg, 1984). Kohlberg defined three levels of moral judgement, each consisting of two stages, although in later work the sixth and highest stage has been excluded because empirically it could not be observed to occur (Colby & Kohlberg, 1987). He proposed that moral judgement universally proceeds successively and irreversibly through these hierarchically ordered stages (Boom, Brugman, & Van der Heijden, 2001). Empirical evidence suggests that in Western countries approximately 10–15% of the people achieve the highest level.

Kohlberg's theory has been extended to include aspects of morality other than abstract moral judgement. One extension concerns the role of moral competence in moral behaviour. Research has drawn attention to the differences between the reasoning underlying abstract hypothetical dilemmas and real-life dilemmas, and has demonstrated that people's moral behaviour need not be consistent with their moral competence (e.g., Haan, 1975; Higgins, Power, & Kohlberg, 1984; Krebs, Denton, & Wark, 1997; Walker, DeVries, & Trevethan, 1987). Rest (1983; Rest, Narvaez, Bebeau, & Thoma, 1999) addressed these discrepancies in a model in which processes other than moral competence, for instance moral perception and moral motivation, also account for moral behaviour. Other approaches have emphasized the contextual specificity of moral judgement. It has been increasingly recognized that moral judgements in real life are not only social in reference (i.e., they refer to human interactions), but also that they mostly arise in social situations and are shared with members of the group to which one belongs. As a consequence, the traditional approach has been supplemented with a more social approach in which subjects are not only asked to reason about real-life dilemmas from their own perspective, which is called practical moral judgement, but also to take the perspective of the majority of the group or context in which they participate. The perception of individuals of these shared values and norms regulating social interactions in moral situations constitute the moral atmosphere or moral culture of a group or institution (Power, Higgins, & Kohlberg, 1989).

Power et al., recognize the difficulty of the task to assess this shared perception from individual reports: ''Individual perceptions are likely to offer only pieces of the whole, colored by individual differences of various kinds'' (Power et al., 1989, p. 109). Yet, these researchers were exclusively interested in constructing ''the whole'' moral atmosphere in school (''moral

culture''). Their characterizations of school moral atmosphere seem to be in part the result of an idealized typing, contrasting just community schools with normal schools. Within-school differences between students in their perception of moral atmosphere were not the object of systematic research. From our viewpoint it is necessary to investigate empirically the contribution of each of the factors involved in school moral atmosphere, like student, class, grade, and school. The moral atmosphere characterized as ''shared'' perception may, after all, explain only a part of the students' perception. Research using other climate measures confirms that such measures can be useful in representing differences at different levels, such as differences between classes and between students within classes (Solomon, Watson, Battistich, Schaps, & Delucchi, 1996, p.730). A more complete picture of the factors contributing to the variance is provided by estimating the variance associated with such levels.

The school is obviously an important social context from childhood to adolescence. The effect of the perceived moral atmosphere in school on real-life moral judgement and behaviour may be particularly strong during adolescence, because during this phase acceptance by the peer group is of utmost importance for individuals (Eccles et al., 1993; Gibbs, Potter, & Goldstein, 1995). Researchers in different traditions have suggested that the moral atmosphere in school can have long-term consequences for students' development. In their just community studies Power et al. (1989) refer to effects on students' career planning, moral judgement development and prosocial behaviour. Goodenow (1993) found positive relations between urban middle-school students' feelings of belonging to school and their academic motivation and effort. Battistich, Solomon, Kim, Watson, and Schaps (1995) found relations between elementary school students' sense of community and their academic attitudes and motives, social and personal attitudes, motives and behaviour, and academic achievement. Hewstone, Jaspars, and Lalljee (1982) reported effects on students' attribution style and identity. According to Battistich et al. (1995, p. 629) differences between these approaches are partly methodological and partly conceptual—the methodological issue pertaining to the focus of measurement: the group or the individual.

Although the moral atmosphere may thus be an important school and/or individual characteristic, larger scale studies on school moral atmosphere as perceived by the students are scarce. Therefore, the universality of the results pertaining to differences in moral atmosphere between schools cannot as yet be considered established. In particular, it should be possible to assess the relative importance of between- and within-school differences in normal secondary schools. In a situation of ''pluralistic ignorance'', which seems to be close to the normal condition of students in regular secondary schools concerning the moral motivation of their peers (Power et al., 1989),

individual characteristics may have a strong influence on adolescents' perception of moral atmosphere. For example, they may attribute negative characteristics to the school atmosphere to justify their own moral behaviour or lack thereof (Gibbs et al., 1995). This viewpoint addresses the perceived moral atmosphere as an instance of the social competence, referring to how well informed the students are about what is going on at school and how capable they are of taking the perspective of other students (cf. Taylor & Walker, 1997).

Studies on a larger scale require instruments for measuring aspects of moral atmosphere in school that are easily administered, such as paper-and-pencil questionnaires. Of course, such instruments should satisfy standard psychometric criteria of reliability. The present study reports reliability estimates of paper-and-pencil questionnaires to measure moral competence and aspects of moral atmosphere in regular secondary schools. Generalizability theory or G-theory (Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is used to estimate the reliabilities of the measures.

The basis of the measurement consists of students' perceptions of other students' moral behaviour, reasoning and opinions. These perceptions need to be combined in some way to obtain measures of moral atmosphere in school. The school score in this study is the average of the students' scores. Although this is a common choice, we wish to emphasize that alternative summaries of students' scores may also yield useful characterizations of moral atmosphere in school (e.g., the highest or lowest score, the mode, or measures of the spread of the within-school score distributions).

METHOD

Sample

Our scales are designed to measure moral atmosphere in secondary schools. A Dutch child leaving primary school at about 12 years of age has a choice of schools of varying educational levels or combinations thereof. The school types selected for the present study represented 66% of the schools and comprised 53% of the students. The school types represent the four educational levels in Dutch secondary schools (i.e., junior vocational; intermediate general; higher general; and university preparatory secondary education). The sample contains small schools as well as larger comprehensive schools, all situated in the western, highly urbanized, part of The Netherlands (a more detailed description of the sample is presented in Høst, Brugman, Tavecchio, & Beem, 1998).

Type 1 schools for junior vocational education provide a four-year programme, which aims at leading the student to further vocational training or education. *Type 2* schools for intermediate general education offer a four-

year general programme leading the student to vocational education or higher secondary education. *Type 3* schools represent the two highest educational levels: higher general education (five years) and university preparatory education (six years). *Type 4* and *Type 5* schools consist of Type 2 and Type 3 educational levels, respectively, within broad-based combined schools. The educational levels are often physically separated and function more or less independently as far as the students are concerned.

Eight schools from each school type participated in this study. Grades 2 and 3 participated in all schools (with mean ages of 13.9 and 15.0 years, respectively). In the two highest educational levels (i.e., higher general and university preparatory education) Grade 4 (with mean age of 15.9 years) also participated in this study. Grade 1 was excluded because at the time of testing students could have only limited experience of the school. Grade 4 was excluded in Type 1 and Type 2 schools because students in those school types take their final exams in Grade 4. From each grade level two classes were randomly sampled. In each class eight students were chosen, if possible, four girls and four boys. In general, the selection was at random, but students could not be forced to participate.

A completely balanced design is very convenient for the estimation of variance components. Therefore, only Grades 2 and 3 were included in the G-study and Type 4 and Type 5 schools are distinguished, although these are actually subdivisions of the same school. The total number of students for the G-study was 1280, with 32 students participating in each school. The missing values on the moral atmosphere items (in total 0.8%) were replaced by the subject's subscale mean or, if a subscale score was missing completely (which happened very rarely), by the item mean of the class.

Instruments

The *School Moral Atmosphere Questionnaire* (SMAQ, Høst et al., 1998) is a multiple-choice instrument. It contains two standardized school dilemmas (one about helping an unpopular classmate, the other about stealing from—and preventing somebody from stealing from—a classmate) to measure the extent that the norms of "helping" and "rejection of stealing" are shared by the students of a school and the collective stage of reasoning concerning these norms. Two content scores were computed: for the norm "helping" the average score over five situations, for "rejection of stealing" the average over three situations. The score on collective stage of reasoning was computed as the mean of five sets of reasons, two from the helping dilemma and three from the stealing dilemma. Henceforth the variables are called *rejection of stealing, helping* and *stage of the norm*.

The SMAQ also contains a questionnaire called "Questions about you and the school" in which students are asked to answer questions from the

perspective of the majority of the students. The questionnaire consists of two scales that refer to "the school as a community" (21 items) and "valuing the school" (13 items), which contain five-point Likert-type items ranging from *absolutely not true* to *absolutely true*. Henceforth these scales are called *community* and *valuation*.

The *Sociomoral Reflection Objective Measure—Short Form*. (SROM-SF, Basinger & Gibbs, 1987) is a paper-and-pencil instrument to measure moral reasoning competence. In total the SROM-SF has 12 sets of four close items and 12 closest items. The SRMS (Sociomoral Reflection Maturity Score) combines the mean close and mean closest score, weighting the first half of the second.

THE GENERALIZABILITY STUDY

Generalizability theory or G-theory recognizes that more sources of variation than enter in the classical test theory definition of reliability can contribute to the observed variance of a measurement. These sources are called facets in G-theory. An observation can be characterized or classified by the combination of the levels of the facets under which the observation may be obtained. Thus facets function in the same way as factors in the terminology of the analysis of variance. The levels of the facets, whose number may be finite or infinite, define a universe of admissible observations.

In a G-study the contributions of facets and possibly their interactions to the total variance are assessed by estimating variance components from a sample of the universe of observations, as familiar from the analysis of variance. The variance components estimated in a G-study can be used to estimate the reliability of the instrument for a particular application, which is called a decision study or D-study. For a D-study a universe score is defined for an object of measurement, corresponding to the true score of classical test theory as the average of a large number of administrations of a test for a subject as the object of measurement. The universe score is usually an average over all levels of facets in the universe of observations and is estimated by a sample from those levels. The reliability or generalizability coefficient is then defined as the ratio of the universe score variance to the observed score variance. It measures how well observed score differences among object of measurements can be used to distinguish the corresponding universe score differences. The reliability generally increases as the number of levels included in the D-study increases, similarly to the increase in classical test theory reliability with an increasing number of items. The number of levels needed for a particular criterion value of the reliability can be estimated from the variance components estimated in the G-study. The universe score of main interest in the present study is the average school

score in the universe of grade levels, classes within grade levels and student within classes.

The linear models for the G-study

We represent the design of the G-study by a linear model. In the linear model we write m for the general mean, a_i for a main effect a whose levels are indexed by i, ab_{ij} for the interaction of the main effects a_i and b_j, and b_{ij} for an effect b_j nested within a_i. The effects *school type, questionnaire items, grade levels, schools within school type, classes, students* and their subscripts are symbolized respectively by t, q, g, s, c, and p, and the subscripts' limits in the sample are symbolized by their respective capitals (e.g., the number of schools sampled within each school type is S and the number of classes within each grade level is C). The main effects t_t and q_q are regarded as fixed and schools are treated as a random effect nested within school type. Although the grade levels included in this study were fixed in advance, results will also be presented for a model in which grade is treated as a random effect nested within schools, since future studies may want to incorporate other grade levels or to sample grade levels at random. Treating grade levels as random can to some extent be justified because the grade levels, instead of being regarded as an experimentally manipulated factor, may be regarded as akin to labels whose content can change from one school to another or from one year to another. In that case, the grade effect resembles a random effect nested within schools rather than a random main effect (if grade levels were treated as a random main effect, the variance component for the main effect of grade would have to be estimated on the basis of only two levels, which does not make sense for a random effect). Therefore, when grade is treated as a random effect, we regard the effect as nested within schools and classes are then nested within grades. This implies that a grade effect that is constant across the population (i.e., grade as main effect) cannot be separated from the effect specific for each school (i.e., the school by grade interaction). The two effects are included in the nested effect. The population of grade levels is regarded as infinite rather than finite because, as we have argued, the content of a certain grade level can vary over schools and time. All other effects that are not fixed are treated as random. Note in particular that the effect of classes, which often have their own unique character, is treated as a random variable, implying that the classes are regarded as a sample from some population. Students are always regarded as a random effect nested within classes.

Variance components are, therefore, estimated for two models, one in which grade is treated as a fixed main effect and one in which grade is treated as a random effect nested within schools. In the first model, the observed

score y_{tqgscp} of student p in class c in grade level g on question q in school s within school type t can be written as:

$$y_{tqgscp} = m + t_t + q_q + g_g + s_{ts} + gs_{gts} + c_{tgsc} + p_{tgscp} + tq_{tg} + tg_{tg} + qg_{qg} + qs_{qts} + qc_{qtgsc} + tqg_{tqg} + qgs_{qgts} + qp_{qtgscp} + e_{tqgscp} \quad (1)$$

Classes are nested within the interaction of school and grade level and qp_{qgtscp} cannot be separated from the replication errors e_{tqgscp} since students answer a question only once. When the grade effect is regarded as nested within schools, the model is:

$$y_{tqgscp} = m + t_t + q_q + s_{ts} + g_{tsg} + c_{tsgc} + p_{tsgcp} + tq_{tq} + qs_{qts} + qg_{qtsg} + qc_{qtsgc} + qp_{qtsgcp} + e_{tqsgcp} \quad (2)$$

As is the custom in G-theory, the terms in the model are defined according to the mixed model described for example by Scheffé (1959, Ch. 8). Thus, the sum over levels of the fixed effects of a model term is zero except if a subscript of a model term refers to a nested effect and the sum is taken over levels within which that effect is nested (e.g., the sum over g of c_{tgsc} in Model 1 is not zero).

An effect that is treated as random with an infinite set of levels in the G-study is sometimes treated as fixed in a D-study, because a fixed subset of the levels is of interest in the D-study (i.e., the universe of generalization is restricted). In contrast, the Models 1 and 2 are population models, which means that the levels of the fixed effects are restricted in the population and all those levels are included in the G-study. In our opinion, the main limitation of this study is that the grade levels were not randomly sampled. Treating grade as a purely random effect can in this case be only approximately valid (perhaps a more faithful model would lie somewhere in between the model with grade as a fixed effect and the model with grade as a random effect). Thus, formally, the results cannot be generalized to other grade levels, and the variance component estimators for Model 2 are probably more biased than for Model 1. Of course, the calculations assume unbiasedness.

Estimation of variance components

The variance components are estimated by the ANOVA method. The expected mean squares corresponding to the effects in the linear models are first expressed as a linear functions of the variance components with a procedure described by, among others, Scheffé (1959, Ch. 8). Next the mean squares are estimated for the models by an univariate analysis of variance, equated to their expected values, and the resulting equations are solved for

the variance components. The expected mean squares corresponding to an effect t are symbolized as $EMS(T)$, its estimates as $MS(T)$, the variance components as σ_t^2 and its estimate as s_t^2.

Generalizability coefficients

Many G-coefficients can be calculated for our design. Although the student scores have a different meaning than the school means, they can clearly be of interest in some studies. The school means G-coefficients are presented for the entire population of schools and for schools of a particular school type, assuming that the within-school-type variances are approximately equal across school types. The student scores G-coefficients are for the entire population and for scores as deviations from the classroom mean. Examples of the calculation of two of these coefficients are given in Appendix 1.

Accuracy of variance component estimators as a function of sample sizes and the variance components

The accuracy of the estimators of the G-coefficients depends on the accuracy of the variance component estimators or, equivalently, of the mean squares. The accuracy of the G-coefficient for school means within a given school type can be evaluated as the accuracy of a ratio of independently distributed mean squares. This is not true for the G-coefficient for school means in the entire population, but these G-coefficients suggest that the accuracies of s_s^2 and s_t^2, the estimators of σ_s^2 and σ_t^2, should be our first concern (see Appendix 1, Table 5).

We calculated for 600 designs, varying the number of levels of factors and the values of the variance components, the accuracy of the estimators of those components for restricted versions of the Models 1 and 2, in which the dependent variable is the sumscore over items and variances of an effect are assumed to be the same across the levels of fixed effects (see Appendix 2 for details). If the Models 1 and 2 are summed over q we get:

$$y_{tqgscp\sim} = m + t_t + g_g + tg_{tg} + s_{ts} + gs_{gts} + c_{gtsc} + p_{tgscp} \qquad (1')$$

for grade g_g a fixed main effect, and

$$y_{tqgscp} = m + t_t + s_{ts} + g_{tsg} + c_{tsgc} + p_{tsgcp} \qquad (2')$$

for grade g_{tsg} a random nested effect. The errors in model $1'$ and $2'$ are denoted by p to emphasize that they include the student effect. The calculations in Appendix 2 demonstrate for σ_t^2 that for our design we can

expect to estimate this component with an error of at most 40% with a probability of at least 0.5. For σ_s^2, we may often get a higher accuracy, making an error of at most 30% with a probability of at least 0.5.

G-coefficients are usually judged as satisfactory for a particular application of the instrument (i.e., a D-study) when they exceed a certain criterion value c. We computed $Prob(r^2 < c \mid \rho^2)$ for $c = .60$ (.05).80, where $\rho^2 = 1 - EMS(C)/EMS(S)$ is the population value of the G-coefficient of school means for the within school type population and $r^2 = 1 - MS(C)/MS(S)$ is its estimate. Figure 1 contains a plot of $Prob(r^2 < c \mid \rho^2)$ for $c = .60$ (.05).80 versus ρ^2 for $S = 8$ and $P = 8$. The values of ρ^2 are obtained from values for the components used in the calculations of the accuracy of the estimates (for details see Appendix 2). The figure shows that $Prob(r^2 < .60 \mid \rho^2 = .70) \approx .20$ and $Prob(r^2 < .80 \mid \rho^2 = .70) \approx .95$. Thus the probability of an error of at most 14% is .75. Figure 1 shows that the error probabilities for our design are quite satisfactory.

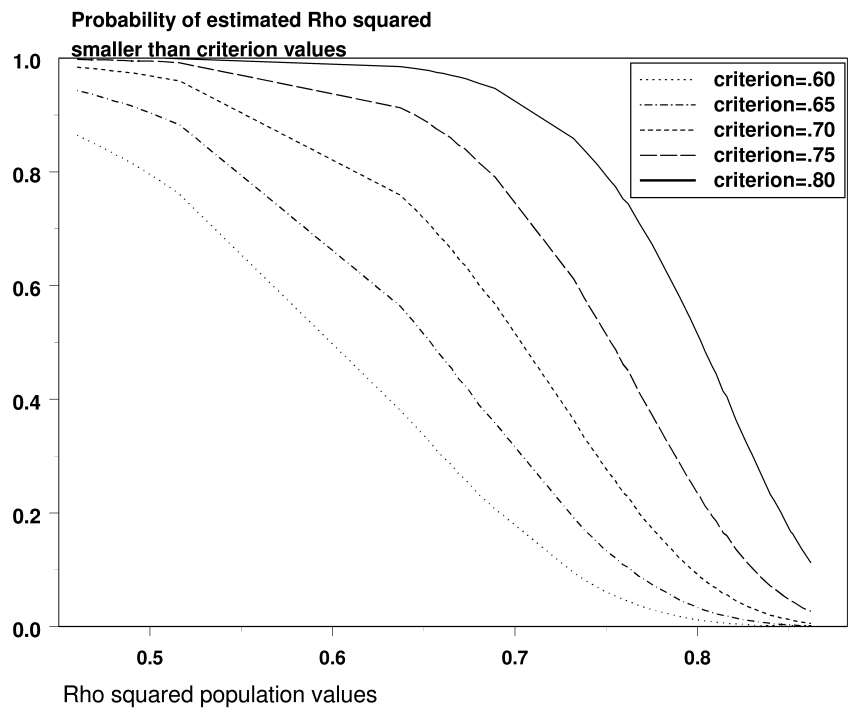

Figure 1. Distribution of estimated rho-squared as a function of population rho-squared for five criterion values.

RESULTS

Table 1 contains component estimates for Model 1 and Table 2 for the Models 1′ and 2′, respectively. Some component estimates are negative. We prefer to present them as such because then the mean squares can be reconstructed from the results and the estimates retain their unbiasedness. Table 1 is mainly of interest because it provides information about s_p^2 independent of the error, and about the question effect and its interactions with other effects. Since for all scales s_p^2 is much larger than the school, grade and class effects, the scales measure to a considerable extent individual variations in the perception of the school environment. The large value of s_q^2 demonstrates substantial variation in the means of the questions. The questions thus seem to span a fairly wide range of the scales they are supposed to measure. For *community* and *moral atmosphere* in Model 1 the estimates of s_{qg}^2, s_{qs}^2, s_{qc}^2 and s_{qgs}^2 are of the same magnitude as the estimates of s_g^2, s_s^2, and s_c^2. This can also be observed for the other scales for at least one of the interactions. To some extent such interactions are conceptually undesirable, although they may also reflect that the frequencies of events

TABLE 1
Variance component estimates for Model 1

	Stage norm	*Rejection stealing*	*Helping*	*Community*	*Valuation*	*Moral atmosphere*
s_t^2	4	20	7	24	224	43
s_q^2	2008	522	72	86,370	84,390	61,620
s_g^2	14	4	0	35	162	43
s_s^2	70	29	8	116	209	95
s_c^2	49	10	0	38	− 9	8
s_p^2	894	164	99	889	1782	612
s_{tq}^2	22	4	4	54	8	53
s_{tg}^2	− 6	3	− 1	− 5	− 2	− 2
s_{qg}^2	16	1	− 1	71	13	48
s_{qs}^2	6	8	1	219	166	147
s_{qc}^2	13	12	− 5	154	133	109
s_{gs}^2	54	0	8	3	100	37
s_{tqg}^2	− 27	2	− 1	1	22	− 1
s_{qgs}^2	111	− 1	10	142	4	77
s_e^2	5781	545	320	12,130	10,690	9127

Note: Component estimates are multiplied by 100.

TABLE 2
Variance component estimates, standard errors, and probability of Ho for Model 1′ and 2′

Scales	*Stage norm* *Model 1′/2′*	*Rejection stealing* *Model 1′/2′*	*Helping* *Model 1′/2′*	*Community* *Model 1′/2′*	*Valuation* *Model 1′/2′*	*Moral atmosphere* *Model 1′/2′*
s_t^2	4.44/4.44	1.78/1.78	1.64/1.64	107.14/107.14	378.80/378.80	1173.76/1173.76
std_T	14.87/14.87	.98/.98	.88/.88	123.76/123.76	158.77/158.77	761.50/761.50
P(H0)	.30/.30	.00/.00	.00/.00	.10/.10	.00/.00	.01/.01
s_g^2	13.88/63.67	.37/.54	− .03/1.89	154.82/150.36	274.09/440.99	1166.32/2127.73
std_g	16.62/43.29	.31/.74	.13/1.06	90.14/159.06	110.58/160.10	540.97/809.57
P(H0)	.05/.09	.01/.26	.39/.04	.00/.20	.00/.00	.00/.00
s_s^2	70.31/38.47	2.57/2.30	1.98/1.04	511.56/436.37	353.27/132.78	2566.25/1502.39
std_s	31.23/35.75	.89/.93	.79/.91	179.88/189.12	115.11/137.07	771.87/857.01
P(H0)	.00/.14	.00/.00	.00/.12	.00/.00	.00/.16	.00/.03
s_c^2	48.80/48.80	.87/.87	.12/.12	169.78/169.78	− 15.59/ − 15.59	211.77/211.77
std_c	37.17/37.17	.76/.76	.84/.84	156.52/156.52	86.66/86.66	462.68/462.65
P(H0)	.06/.06	.09/.09	.42/.42	.11/.11	.55/.55	.30/.30
s_{tg}^2	− 5.61/—	.24/—	− .17/—	− 22.66/—	− 3.19/—	− 52.43/—
std_{tg}	12.03/—	.35/—	.31/—	29.50/—	39.47/—	195.12/—
P(H0)	.57/—	14/—	.60/—	.64/—	.45/—	.52/—
s_{gs}^2	54.27/—	− .03/—	2.06/—	13.67/—	169.45/—	1003.35/—
std_{gs}	43.64/—	.66/—	1.16/—	139.62/—	109.78/—	610.10/—
P(H0)	.08/—	.50/—	.02/—	.45/—	.04/—	.03/—
s_p^2	1471.61/1471.61/—	31.12/31.12/—	40.82/40.82	6466.37/6466.37/—	4401.25/4401.25/—	21295.4/21295.4/—
std_p	62.19/62.13	1.31/1.31	1.72/1.72	273.25/273.01	185.99/185.82	899.09/899.09

Note: Component estimates are multiplied by 100.

in one environment are not in direct proportion to those in another environment. The interactions do not decrease the reliabilities because the questions effect is a fixed effect and the interactions can be estimated (if such effects also exist at the student level, where the interaction cannot be estimated, then the true reliabilities are probably higher than their estimates). Moreover, if the questions effect is regarded as a random effect, which would be appropriate when, for example, different subsets of questions are randomly selected for different schools, then the interactions decrease the reliabilities. Similar estimates are obtained in Model 2, except that s_g^2 is larger in Model 2, because it includes the *gs* interaction.

Table 2 contains the component estimates for Models 1′ and 2′, the unbiased estimates of their standard errors, and P(H0), the probability of the null hypothesis that the component is zero. The form of the statistics for the latter test can be inferred from the expected mean squares, which show that for each component in the restricted model a ratio of two mean squares can be formed which differ by that component only. These ratios have an F-distribution under the null hypothesis if the numerator is the mean square with the component.

The conventional .05 significance level is reached for s_t^2 for three scales and for s_s^2 for all scales in Model 1′ and for three scales in Model 2′. Except that s_p^2, s_s^2 is often the largest component, or nearly so, in Model 1′, whereas in Model 2′ s_g^2 is mostly the largest. Grade two scores are in general higher than grade three scores, which explains the grade effect in Model 1′. This effect is reversed for approximately 25% of the schools, which accounts for the grade by school interaction. It should also be observed that s_s^2 is generally larger than s_t^2, especially for *community*, which is perhaps due to *community* hardly having intellectual aspects. The class effect is nowhere significant and is often smaller than the grade effect or the grade by school effect.

The difference between Models 1′ and 2′ is due to treating grade as a random effect nested within schools in Model 2′. Variation among schools is therefore tested against variation among grades, which includes s_g^2 and s_{gs}^2 of Model 1′, instead of variation among classes. Viewed in another way, school means in Model 2′ are less certain estimates than in Model 1′ because in Model 2′ the average is taken over a random sample, not the complete set of levels, of an additional random effect—grades. With two exceptions, the grade effect is even larger than the school effect in Model 2′. As was also evident in Models 1 and 2, the error component, which here includes the student component, is the largest.

G-coefficients for the Models 1 and 2 are presented in Tables 3 and 4, respectively. The G-coefficient for *srom* is shown for comparison, because it is a strictly individual measure of moral reasoning competence. The *srom* score may be related to school type, as school type is related to the

TABLE 3
G-coefficients Model 1

	School means; S = 8, G = 2, C = 2 scores						*Pupil*	
	WT			*EP*				
Scale	*P = 8*	*P = 16*	*P = 24*	*P = 8*	*P = 16*	*P = 24*	*EP*	*WC*
Stage norm	.55	.67	.72	.56	.68	.73	.64	.61
Rejection stealing	.68	.79	.83	.77	.85	.88	.55	.47
Helping	.60	.75	.81	.72	.83	.88	.64	.61
Community	.68	.78	.82	.71	.81	.84	.65	.61
Valuation	.73	.84	.89	.83	.91	.94	.74	.68
Moral atmosphere	.78	.87	.90	.83	.90	.93	.82	.78
SROM—SF	.38	.53	.60	.77	.86	.89	.71	.66

Note: WT is within-school type; EP is entire population; WC is within classes.

TABLE 4
G-coefficients school means Model 2

	S = 8, C = 2, P = 8/16					
	WT			*EP*		
Scale	*G = 2*	*G = 3*	*G = 4*	*G = 2*	*G = 3*	*G = 4*
Stage norm	.30/.36	.39/.46	.46/.53	.32/.39	.41/.48	.48/.56
Rejection stealing	.61/.70	.70/.78	.76/.83	.72/.79	.79/.85	.84/.88
Helping	.32/.39	.41/.49	.48/.56	.51/.59	.61/.69	.68/.74
Community	.58/.67	.67/.75	.73/.80	.62/.70	.71/.78	.77/.83
Valuation	.27/.32	.36/.41	.43/.48	.55/.60	.65/.70	.71/.75
Moral atmosphere	.46/.51	.56/.61	.63/.67	.58/.63	.67/.72	.73/.77
SROM – SF	.07/.09	.10/.13	.13/.17	.66/.73	.74/.80	.79/.84

Note: WT is within-school type; EP is entire population.

educational level. In Table 3 the number of students is varied and the number of grade levels and classes is kept at our sample values. We think that in general a G-coefficient should be at least as large as .65. The G-coefficients for school means are fairly satisfactory, but the G-coefficient of *stage of the norm* and *helping* for the within school type population satisfy the criterion only if $P > 8$. The G-coefficients for student scores in the entire population (i.e., the conventional reliability coefficient) are mostly lower than the corresponding school mean G-coefficients, although the difference is negligible for *moral atmosphere* (Brennan, 1995, discusses the relationships between group and individual level G-coefficients). The G-coefficients for Model 2 in Table 4 are much lower. In Table 4 both the number of grades and the number of students are varied. Only *rejection of stealing, community*

and, through them, *moral atmosphere* reach acceptable values of the G-coefficients. For *moral atmosphere*, quite a large number of observations appears to be required, especially for the within school type population.

As an alternative to the simple total scale score, the subscales may be transformed so that their minimum and maximum attainable scores are the same by subtracting the minimum attainable score and then dividing by the maximum attainable score. The *moral atmosphere* G-coefficients with these weights are slightly lower in Model 1 but substantially lower in Model 2, because the relative weights for the less reliable subscales become higher. Sum scores for several combinations of three subscales yield G-coefficients nearly the same as the total scale G-coefficients (e.g., *rejection of stealing, valuation, community*; *helping, valuation, community*).

SUMMARY AND DISCUSSION

In this study we have presented reliability estimates of moral atmosphere in school as a function of the number of grade levels, classes and pupils in the sample. Moral atmosphere in school is defined here as the average perception of students of their schoolmates' moral behaviour, reasoning and opinions. The reliability estimates were calculated from variance component estimates of effects in two models. In Model 1 grade level was treated as a fixed effect, in Model 2 as a random effect. We showed that our reliability estimates are probably fairly accurate, whereas the accuracy of the school type and school variance component estimates is probably only moderate.

The results demonstrate that differences between schools in moral atmosphere in school can be measured reliably, but a considerable number of observations may be needed, especially if grade level is treated as a random effect. The largest proportion of the total score variance is at the student level. The class effect is often smaller than the school and grade effect. Perceptions of the moral atmosphere are, then, influenced by the classroom or classmates to a lesser extent than one might expect.

Although it is very common to find the largest variance at the student level, it is a discomforting feature for a moral atmosphere measurement instrument that the student and error variance components are as large as they are here. We mention two of a number of factors that may be responsible for the large variance at the student level. Perhaps the most unpleasant reason for this result (as opposed to being due to response tendencies) would be that students intrinsically perceive the majority perspective of the same situations very differently. Students may also base their judgement on different situations or attach different weights to the same situations. Because of differences in sociometric status students may have a different access to relevant situations in school, be more or less

acquainted with the perspectives of other students, and be more or less able to understand the impact of these situations on other students (cf. Taylor & Walker, 1997). As such perceived moral atmosphere would be regarded as an instance of social competence, but the size of the class component for *rejection of stealing* and *community*, although relatively small, suggests that such effects are influenced by local effects.

The situations on which judgements are based can collectively be regarded as a subset of all the relevant situations in the school. If the subset is representative of all the relevant situations, then the instrument nevertheless provides a reasonable basis for comparing schools.

The relative importance of these factors can be investigated by having students discuss their answers in a group. We expect that such an approach would reduce the pluralistic ignorance in school among students about each others' opinions. Any educational strategy for such discussions has the additional advantage of being useful already at the individual level, discussion group level, and class level, i.e., the research on school level can be replaced by this more flexible strategy. Because of the substantive relationship between moral atmosphere and undesirable behaviour, such an approach may nevertheless have beneficial educational effects of the students involved. However, when such an approach is feasible at the school level, it may also contribute to an understanding of the relative contribution of student characteristics, the teachers and the school policy to the moral atmosphere.

REFERENCES

Basinger, K. S., & Gibbs, J. C. (1987). Validation of the sociomoral reflection objective measure—short form. *Psychological Reports*, *61*, 139–146.

Battistich, V., Solomon, D., Kim, D, Watson, M., & Schaps, E. (1995). Schools as communities, poverty levels of student populations, and performance: A multilevel analysis. *American Educational Research Journal*, *32*, 627–658.

Boom, J., Brugman, D., & Van der Heyden, P. G. M. (2001). Hierarchical structure of moral stages assessed by a sorting task. *Child Development*, *72*, 535–548.

Brennan, R. L. (1992). *Elements of generalizability theory* (Rev. ed.). Iowa City, IA: American College Testing.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement*, *32*, 385–396.

Colby, A., & Kohlberg, L. (1987). *The measurement of moral judgement. Vol. 1: Theoretical foundations and research validation*. Cambridge: Cambridge University Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, H. (1972). *The dependability of behavioural measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Mac Iver, D. (1993). Development during adolescence. The impact of stage–environment fit on young adolescents' experiences in schools and families. *American Psychologist*, *48*, 90–101.

Gibbs, J. C., Potter, G. B., & Goldstein, A. P. (1995). *The EQUIP Program. Teaching youth to think and act responsibly through a peer-helping approach*. Champaign, IL: Research Press.

Goodenow, C. (1993). Classroom belonging among early adolescent students. Relationships to motivation and achievement. *Journal of Early Adolescence*, *13*, 21–43.

Haan, N. (1975). Hypothetical and actual moral reasoning in a situation of civil disobedience. *Journal of Personality and Social Psychology*, *32*, 255–270.

Hewstone, M., Jaspars, J., & Lalljee, M. (1982). Social representations, social attribution and social identity: The intergroup images of "public" and "comprehensive" schoolboys. *European Journal of Social Psychology*, *12*, 241–269,

Higgins, A., Power, C., & Kohlberg, L. (1984). The relationship of moral atmosphere to judgements of responsibility. In W. L. Kurtines & J. L. Gewirtz (Eds.), *Morality, moral behaviour and moral development* (pp. 74–106). New York: Wiley.

Høst, K., Brugman, D., Tavecchio, L. W. C., & Beem, A. L. (1998). Students' perception of the moral atmosphere in secondary schools and the relationship between moral competence and moral atmosphere. *Journal of Moral Education*, *27*(1), 47–71.

Kohlberg, L. (1981). Exploring the moral atmosphere of institutions: A bridge between moral judgement and moral action. In L. Kohlberg, *The meaning and measurement of moral development* (pp. 35–53). Worcester, MA: Clark University Press.

Kohlberg, L. (1984). *The psychology of moral development*. San Francisco: Harper & Row.

Kohlberg, L. (1985). The just community approach to moral education in theory and practice. In M. W. Berkowitz & F. Oser (Eds.), *Moral education: Theory and application* (pp. 27–87). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Krebs, D. L., Denton, K., & Wark, G. (1997). The forms and functions of real-life moral decision-making. *Journal of Moral Education*, *26*, 131–143.

Power, C., Higgins, A., & Kohlberg, L. (1989). *Lawrence Kohlberg's approach to moral education*. New York: Columbia University Press.

Rest, J. R. (1983). Morality. In J. H. Flavell & E. Markman (Eds.), *Manual of child psychology: Vol. 3. Cognitive development* (pp. 555–629). New York: Wiley.

Rest, J., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999). *Post-conventional moral thinking. A neo-Kohlbergian approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Scheffé, H. (1959). *The analysis of variances*. New York: Wiley.

Searle, S. R. (1971). *Linear models*. New York: Wiley.

Solomon, D., Watson, M., Battistich, V., Schaps, E., & Delucchi, K. (1996). Creating classrooms that students experience as communities. *American Journal of Community Psychology*, *24*, 719–748.

Solomon, D., Battistich, V., & Hom, A. (1997). Teacher beliefs and practices in schools serving communities that differ in socioeconomic level. *The Journal of Experimental Education*, *64*, 327–347.

Taylor, J. H., & Walker, L. J. (1997). Moral climate and the development of moral reasoning: The effects of dyadic discussions between young offenders. *Journal of Moral Education*, *26*, 21–43.

Walker, L. J., De Vries, B. J., & Trevethan, S. D. (1987). Moral stages and moral orientations in real-life and hypothetical dilemmas. *Child Development*, *58*, 842–858.

Manuscript received 10 June 2003
Manuscript accepted 23 March 2004

APPENDIX 1

Generalizability coefficients are defined as the ratio of the universe score variance to the expected observed score variance for a particular object of measurement and score of interest. The calculations of two of the G-coefficients in Table 5 are presented here for Model 1. The first is the G-coefficient for school means in the entire population. The universe score is $m + t_t + s_{ts}$, its population mean is m and thus the universe score variance is $\Sigma_t E_s(t_t + s_{ts})^2/T = (T-1)\sigma_t^2/T + \sigma_s^2$, where we follow the custom of using the same symbol, σ^2, for the variance component and for quadratic forms such as $\Sigma_t t_t^2/(T-1)$. The observed school mean is

TABLE 5
G-coefficients of school means and within-classroom pupil scores for two target populations

Model 1

Score: School means; Target population: entire population

$$[(T-1)\sigma_t^2/T + \sigma_s^2]/(T-1)\sigma_t^2/T + \sigma_s^2 + \sigma_c^2/GC + \sigma_p^2/GCP = \sigma_e^2/QGCP$$

Score: School means; Target population; within-school type

$$\sigma_s^2/(\sigma_s^2 + \sigma_c^2/GC + \sigma_p^2/GCP + \sigma_e^2/QGCP)$$

Score: Pupil means; Target population: entire population

$$[(T-1)\sigma_t^2/T + (G-1)\sigma_g^2/G + (T-1)(G-1)\sigma_{tg}^2/TG + \sigma_s^2 +$$
$$(G-1)\sigma_{gs}^2/G + \sigma_c^2 + \sigma_p^2]/[(T-1)\sigma_t^2/T + (G-1)\sigma_g^2/G +$$
$$(T-1)(G-1)\sigma_{tg}^2/TG + \sigma_s^2 + (G-1)\sigma_{gs}^2/G + \sigma_c^2 + \sigma_p^2 + \sigma_e^2/Q]$$

Score: Pupil means; Target population: within classrooms

$$\sigma_p^2/(\sigma_p^2 + \sigma_e^2/Q)$$

Model 2

Score: School means, Target population: entire population

$$[(T-1)\sigma_t^2/T + \sigma_s^2]/[(T-1)\sigma_t^2/T + \sigma_s^2 + \sigma_g^2/G + \sigma_c^2/GC + \sigma_p^2/GCP + \sigma_e^2/QGCP]$$

Score: School means; Target population: within-school type

$$\sigma_s^2/(\sigma_s^2 + \sigma_g^2/G + \sigma_c^2/GC + \sigma_p^2/GCP + \sigma_e^2/QGCP)$$

Score: Pupil means; Target population: entire population

$$[(T-1)\sigma_t^2/T + \sigma_s^2 + \sigma_g^2 + \sigma_c^2 + \sigma_p^2]/[(T-1)\sigma_t^2/T + \sigma_s^2 + \sigma_g^2 + \sigma_c^2 + \sigma_p^2 + \sigma_e^2/Q]$$

Score: Pupil means; Target population: within classrooms

$$\sigma_p^2/(\sigma_p^2 + \sigma_e^2/Q)$$

$y_{ts\sim} = m + t_t + s_{ts} + c_{ts\sim} + p_{ts\sim} + e_{ts\sim}$, where a tilde signifies the arithmetic average over the sample values for the subscripts that are replaced by the tilde (e.g., $y_{ts\sim} = \Sigma_{qgcp} y_{tqgscp}/QGCP$, and $c_{ts\sim} = \Sigma_{gc} c_{gtsc}/GC$). Its population mean is m and the variance becomes $\Sigma_t E_s (t_t + s_{ts} + c_{ts\sim} + p_{ts\sim} + e_{ts\sim})^2/T = (T-1)\sigma_t^2/T + \sigma_s^2 + \sigma_c^2/GC + \sigma_p^2/GCP + \sigma_e^2/QGCP$.

As a second example, the G-coefficient for the student scores as deviation from the population mean is calculated. The observed student mean score in Model 1 is $\Sigma_q y_{tgscp\sim} = m + t_t + g_g + s_{ts} + gs_{gts} + c_{tgsc} + p_{tgscp} + tg_{tg} + e_{tgscp\sim}$ and the expected value over the entire population is m. The variance is, therefore, $\Sigma_{tg} E_{scpe} (t_t + g_g + s_{ts} + gs_{gts} + c_{tgsc} + p_{tgscp} + tg_{tg} + e_{tgscp\sim})^2/TG = (T-1)\sigma_t^2/T + (G-1)\sigma_g^2/G + (T-1)(G-1)\sigma_{tg}^2/TG + \sigma_s^2 + (G-1)\sigma_{gs}^2/G + \sigma_c^2 + \sigma_p^2 + \sigma_e^2/Q$, since the average over g of the covariances $E_s s_{ts} g s_{gts}$ is zero as a consequence of the model definition. The universe score variance is the expected observed score variance minus σ_e^2/Q, because the universe score is the expectation of the observed score over repeated administration of the test.

Table 5 contains expressions in terms of variance components for G-coefficients for school means, which are the primary concern of this study, and for student scores in Models 1 and 2.

APPENDIX 2

Accuracies were assessed by computing probabilities for intervals of the form $(1-h)\sigma_s^2 < s_s^2 < (1+h)\sigma_s^2$, for $h = 0.2$ (0.1) 0.6, for various sample sizes and variance component population values. The sample sizes for school types, grades and classes were fixed at our sample values (i.e., 5, 2 and 2, respectively). The number of schools S for each school type and the number of students p. in each class were set to 6 (2) 12 and 6, 8, 10, respectively. The variance components were scaled to sum to one. The values of σ_s^2 and σ_t^2 were set to 0.03 (0.03) 0.15. This range for σ_s^2 is fairly typical for school effects and σ_t^2 was presumed to be approximately in the same range. The value of σ_p^2 was set to $b(1 - \sigma_s^2 - \sigma_t^2)$ for $b = 0.8$ or 0.9, since variation at the student level is typically considerably larger than variation at the school level. The values for the remaining components were all set to $(1 - \sigma_s^2 - \sigma_t^2 - \sigma_p^2)$ divided by the number of remaining components. This resulted in 600 combinations of different sample sizes and component values.

The distribution of component estimators was approximated by the Satterthwaite approximation and by the normal distribution, which is valid in large samples. Only the results of the latter are presented, because the results

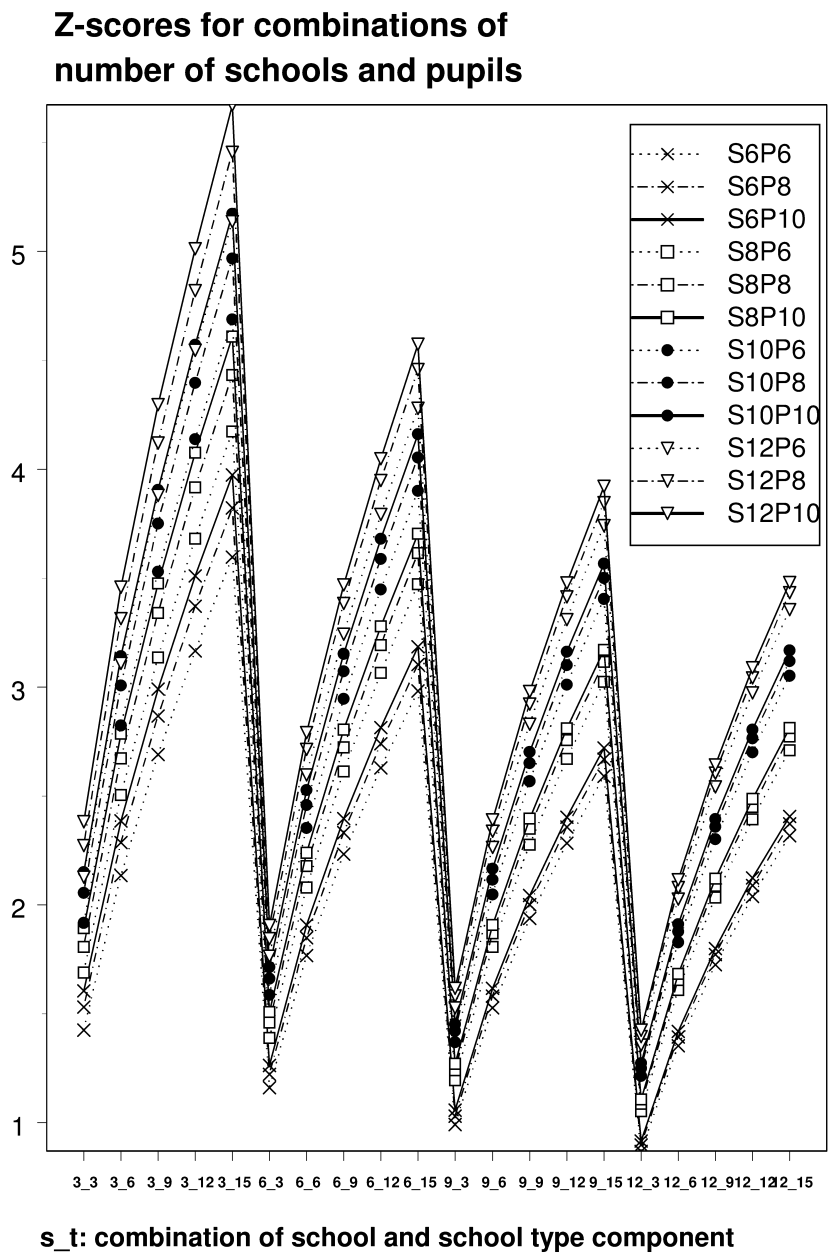

Figure 2. *Z*-scores for σ_t^2 as a function of component values and number of factor levels.

of the two approximations were generally very similar for the intervals considered here.

Applying the normal approximation to $MS(S) - MS(C)$ and $MS(T) - MS(S)$, we have approximately a standard normal distribution for $k_s(s_s^2 - \sigma_s^2)/\{var[MS(S) + var[MS(C)]\}^{1/2}$ and $k_t(s_t^2 - \sigma_t^2)/\{var[MS(T)] + var[MS(S)]\}^{1/2}$, where $k_s = GCP$ and $k_t = SGCP$ (note that the number of school types is fixed and cannot become large). The variances of the mean squares can be calculated by applying a basic theorem on the distribution of quadratic forms of normally distributed variables (e.g., Searle, 1971).

If $(s_s^2 - \sigma_s^2)/q_s$ for $q_s = \{var[MS(S)] + var[MS(C)]\}^{1/2}/k_s$ has a standard normal distribution, then $Prob(s_s^2 < (1 + h)\sigma_s^2)$ can be computed from $Prob(z < h\sigma_s^2/q_s)$, where z has a standard normal distribution. If the probability of an error of at most 30% ($h = .3$) is desired to be at least .5,

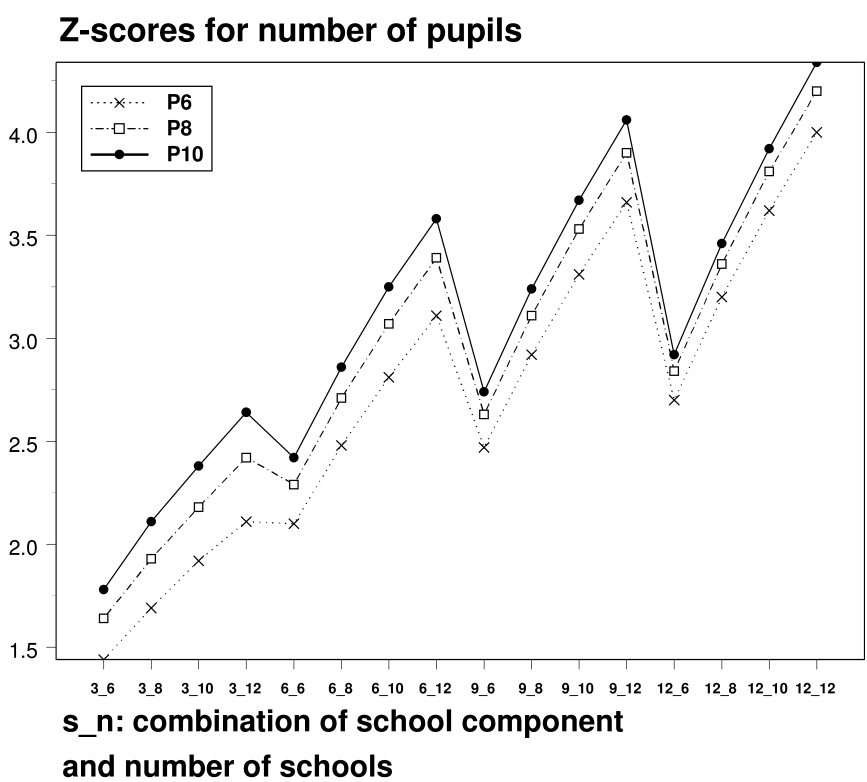

Figure 3. Z-scores for σ_s^2 as a function of component values and number of factor levels.

then σ_s^2/q_s should be at least .68/.3 = 2.27, as $Prob(0 < z < .68) \approx .25$. Similarly, an error of at most 40% with a probability of at least .5 corresponds to $z = 1.7$. We will use 2.27 and 1.7 as guiding figures.

Only the results for Model 1′ will be presented. Figures 2 and 3 depict the relationship between σ_t^2/q_t and combinations of σ_t^2, σ_s^2, S and P, and between σ_s^2/q_s and combinations of σ_s^2, S and P, respectively. In an ANOVA those factors and some of their interactions accounted for .993 and .997 of the variance of σ_t^2/q_t and σ_s^2/q_s, respectively. The first number of the x-axes labels signifies the value of σ_s^2, the second number signifies σ_t^2 in Figure 2 and S in Figure 3. The y-axes represent the means of σ_t^2/q_t in Figure 2 and of σ_s^2/q_s in Figure 3. The means are depicted for each combination of S and P in Figure 2, in which for example $S8P6$ signifies the combination of eight schools and six students in each class, and for each P in Figure 3 (differences between the mean and minimum values were minor). The values for $\sigma_s^2 = .15$ were nearly identical to those of $\sigma_s^2 = .12$ and are therefore not shown.

In Figure 2 the largest effect is the increase of σ_t^2/q_t as the difference $\sigma_t^2 - \sigma_s^2$ increases. Hence, as σ_s^2 increases along the x-axis, σ_t^2/q_t generally decreases. For a given value of $\sigma_t^2 - \sigma_s^2$, σ_t^2/q_t is nearly constant. The number of schools has a stronger effect than the number of students. The effect of the latter becomes small for high values of σ_s^2, as is evident from the clustering of the lines at the right of Figure 2. It is therefore more effective to increase the number of schools than to increase the number of students. Of course, a larger number of schools will in general require more resources and be more difficult to recruit. If eight schools and at least six students in each class are sampled, then σ_t^2/q_t exceeds 2.27 for $\sigma_s^2 > .03$ and $\sigma_t^2 > \sigma_s^2$. But if $\sigma_t^2 - \sigma_s^2 < -.03$, then as many as 12 schools are barely sufficient. Thus, a value of 1.7 seems a more realistic criterion when $\sigma_t^2 - \sigma_s^2$ is negative. Our sample of eight schools approximately satisfies this criterion over a reasonable range of σ_s^2, but twelve schools are clearly a better choice when enough resources are available.

In Figure 3 the largest effect is the increase of σ_s^2/q_s with the increase of σ_s^2 and S. If $\sigma_s^2 > .06$, then σ_s^2/q_s is always larger than 2.27, but if $\sigma_s^2 \leqslant .06$, then σ_s^2/q_s is larger than 2.27 only if $S \geqslant 10$ and $p = 12$. The values $S = 8$ and $p = 8$ seem to strike a reasonable balance in that the criterion is satisfied or nearly so if $\sigma_s^2 > .03$. A comparison of Figures 2 and 3 shows that larger values of σ_s^2 have opposite effects in these Figures.